ED 092 597 TM 003 709

AUTHOR	Convey, John J.
TITLE	Some Methodological Considerations for School Effectiveness Studies.
PUB DATE	Oct 73
NOTE	69p.; Paper presented at the Annual Meeting of the Florida Educational Research Association (Tallahassee, Florida, January 1974)
EDRS PRICE	MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS	*Academic Achievement; Measurement Techniques; *Models; Predictor Variables; Productivity; *Research Methodology; *Schools; *Statistical Analysis
IDENTIFIERS	Dyer Model; Measurement of Change; Production Process Model; *School Effectiveness

ABSTRACT

Some of the methodological considerations in school effectiveness studies are outlined and a state of the art presented. Two general theoretical models are given which provide the researcher with an overall strategy for handling such a study of school effectiveness: the Dyer Model and Production Process Model. Six statistical models which provide possible methods for the computation of effectiveness indices are proposed and critiqued: (1) analysis of covariance (ANCOVA), (2) nonstandard ANCOVA, (3) corrected nonstandard ANCOVA, (4) mean differences scores, (5) individual regression residuals, and (6) school regression residuals. Finally, several other technical considerations involving sources of error, identification of predictors, choice of input and output, unit of analysis, type of samples, and the kind of analysis to be performed are briefly discussed. Major emphasis is on models used to rank schools in terms of effectiveness. (Author/RC)

SOME METHODOLOGICAL CONSIDERATIONS
FOR SCHOOL EFFECTIVENESS STUDIES

A paper presented to the
Evaluation and Research Design Program
of the College of Education
of Florida State University

John J. Convey
Florida State University
October, 1973

* Presented while the author was a graduate student

INTRODUCTION

The determination of which schools are operating most effectively in terms of student development has always been of interest to educators and the general pubiic. In recent years, the question has been receiving more attention due to the increasing demand for accountability. Spurred by the results of the controversial Equality of Educational Opportunity study (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, & York, 1966), many school effect studies have been conducted over the past several years. It appears that the issue of determining school effectiveness will continue to be an important one in the future.

At the present time, there appears to be a need to provide for researchers some guidelines as to the state of the research in school effectiveness studies. The purpose of this paper is to present some of the methodological considerations that must be made in such a study.

First, some assumptions which seem to be necessary are briefly discussed. Two general theoretical models are then presented to offer the researcher an overall strategy for handling such a study. Next, six statistical models which provide possible methods for the computation of effectiveness indices are proposed and critiqued. Finally, several other technical considerations involving sources of error, identification of predictors, choice of input and output, unit of analysis, type of samples, and the kind of analysis to be performed are briefly discussed.

Since the paper is intended merely to acquaint the school-effectiveness researcher with some of the problems he will encounter, discussion of each of these issues is kept to a minimum. An explicit and complete "how to" guide is beyond the scope of this paper. In addition, statistical considerations are kept to a minimum, although the researcher should realize that such studies will involve a good deal of statistical work. It is assumed that the reader is familiar with the fundamentals of the analysis of covariance and multiple linear regression.

Before proceeding further, a distinction should be made between two related, but different, aspects of trying to determine which schools are more effective than others. Some researchers are interested in the question: Which school characteristics are the best predictors of student progress? The related question which immediately follows is: Given these predictors, which schools are more effective than others? The former is essentially an hypothesis testing problem in that the researcher hypothesizes that a particular school variable does influence student progress, and then proceeds to test that hypothesis. On the other hand, the latter is essentially a prediction problem since it involves comparisons on the dependent variable. That both are related is clear. Effective predictors of student progress must be identified, so that they can be used in the determination of effective schools. If the attempt to make this determination is done in the absence of effective predictors, then the results of the analysis may lead to erroneous conclusions about which schools are more effective. The major thrust of this paper will concentrate on models used to rank schools in terms of effectiveness.

SOME ASSUMPTIONS

Prior to presenting the general theoretical models, the statistical models and the other technical considerations, several assumptions which the researcher should realize that he must be prepared to make in conducting a school effectiveness study should be made expl.cit.

The first assumption is that there are real differences in effectiveness from school to school along at least one dimension. Furthermore, that dimension can be identified. The researcher must assume that the dimension chosen on which to compare schools in terms of effectiveness is one along which the schools really differ. When a difference is observed after application of one of the models, at least part of the difference is due to differential effectiveness and not merely to artifacts of the statistical method emoloyed. So, there are real differences present, and the model is helping to make them explicit.

The second ass umption is that measurable output variables can be determined which will adequately represent a dimension along which schools are differentially effective. For example, suppose that math achievement is one such dimension. Any output variable that is used as a measure of math achievement is in reality only a substitute for math achievement. It is assumed that the score on the math test adequately represents the dimension of math achievement, and that schools can be ranked on this basis.

The third assumption is that measurable differences in student outcome are attributable to measurable differences in school variables which can be manipulated. This appears to be a basic assumption of such studies. If school variables which influence change in student outcomes cannot be identified, then there may not be really such a thing as a school effect. If schools turn out to be differentially effective, not because of what they are or what they do, but merely because of the type of student they have or the neighborhood in which they are located, then school effectiveness may be a misnomer. Furthermore, if the effective school variable cannot be manipulated, school effects studies become an exercise in frustration. However, this latter consideration may concern the administrator more than the researcher.

The fourth assumption is that, given the goals or objectives against which the schools are to be compared in terms of effectiveness, all schools considered in the study are trying to maximize the same group of goals or objectives for all students. There is certainly a problem with this assumption, and the researcher should be aware of it. Priorities do vary among schools, and to the extent that emphasis of basic goals is different, any attempt to compare schools on these will be inadequate. This realization should encourage the researcher to select those dimensions of comparisons which are time-honored in most schools. For example, most schools have as an objective to increase the basic reading ard math skills of their students. Emphasis here may vary, but probably not to the extent that it would in areas such as moral development or physical fitness. The seriousness of failing to meet this assumption depends on the purpose of the study. In studies designed to identify effective predictors, variables may emerge as important only because of the different emphasis given to the goals represeuted by the dependent measures. This may introduce or perpetuate the use of improper predictors. On the other hand, in studies designed to identify effective schools, this failure is likely to result in the obvious finding that schools which do not place much emphasis on the particular goals show up to be less effective.

THE GENERAL MODELS

Standardized tests of academic achievement have long been used for evaluating the effectiveness of an individual school or school system. Typically, the approach has been to compare the mean performance of the school or system with some local or national norm, and to assume that the discrepancy between the two measures constitutes an indication of the effectiveness of the school or system. There are many problems associated with this method, not the least of which is that only output is considered, and probably only achievement output, and such variables as entering student characteristics and what goes on in the school are completely neglected.

In the following sections, two general theoretical models which treat output as a function of input and other variables are reviewed. These models are rather similar and differ only in the way they conceive the relationships. They are offered to the researcher as general plans for determining school effectiveness. Later, specific statistical models which may serve as tools within the context of the two general models will be discussed.

The Dyer Model

An intuitively pleasing theoretical model for handling the determination of school effectiveness has been offered by Henry Dyer. Dyer (1966), in outlining a technique for the evaluation of school systems for Pennsylvania, suggested that a discrepancy measure of school system effectiveness might be based on the deviation between the mean achievement scores actually found at any grade level, and the mean achievement scores predicted from measures of previous student achievement and the hard-to-change conditions that presumably affect the learning process. Dyer has since elaborated on this concept (1966, 1967, 1969, 1970a, 1970b, 1972a, 1972b, 1972c), and the model has become known as the Dyer Accountability Model or the Student Change Model.

Four groups of variables are considered by the model:

1. Output - the performance of students at the end of a particular phase of schooling. Output consists of all the measured characteristics of students as they finish a particular phase of their schooling: command of basic skills, state of health, appreciation of their roles as citizens, attitudes, interests, achievement in various areas, aspirations, social behavior, moral development, and so on.

2. Input - the performance of students at the beginning of some particular phase of schooling. Input basically consists of initial measures along the same dimemsions as the output variables.

3. Surrounding Conditions - what went on outside the system that may have helped or hindered the development of the students. Surrounding conditions can be divided into home variables, school variables, and community variables. Some of these will be classified as easy to change,

while others are hard to change. This distinction is important in Dyer's application of the Model, since the hard-to-change conditions appear as predictors in the regression system, while the easy-to-change conditions are used to discover ways to improve the system.

4. Process Variables - what went on inside the system that may have been productive or counterproductive. These are closely related to surrounding condition variables and are frequently confused with such. It is important to distinguish these in Dyer's use of the Model. For example, the number of books in the school library is a measure of school conditions, while the rate at which the books are actually used is a measure of process. Similarly, the math teachers' backgrounds and experiences are measures of school conditions, while the number of creative projects they stimulate in the students is a measure of process.

A scheme needs to be developed to rank schools on effectiveness based on relationships between these variables. Later on in this paper several such schemes are discussed. Dyer conceived the model to be used in the following way. The regression of output on input and the hard-to-change surrounding condition variables is obtained. Residuals are calculated by taking the difference between the observed output (0) and the predicted output ($\hat{0}$). An index (I) is then computed as follows.

$$I = \frac{0 - \hat{0}}{\frac{SD}{\sqrt{n}}}$$

where, O is the output mean for a school

$\hat{O}$ is the predicted output mean for the school

$\overline{SD}$ is the average within-school standard deviation on the output

$\bar{n}$ is the average number of students per school on the output.

Performance indices (PI) are defined as follows: (Dyer, Linn, & Patton, 1967)

$$\begin{array}{rcccll} & & I & < & -1.5, & PI = 1 \\ -1.5 & \leq & I & < & -.5, & PI = 2 \\ -.5 & \leq & I & < & .5, & PI = 3 \\ .5 & \leq & I & \leq & 1.5, & PI = 4 \\ 1.5 & < & I & & , & PI = 5. \end{array}$$

The Pl's are then used to identify schools that seem to be performing either above expectation or below expectation with respect to a particular class of educational outcomes. After such schools have been identified, the strategy is to investigate the easy-to-change surrounding condition variables and the process variables in order to try to account for the differential performance.

Dyer (1970a) provides the following hypothetical example for a school system. Suppose performance indicators are calculated for four levels of a system using five different output measures. These are summarized in Table 1.

The system seems to be doing better in some areas of student development than in others. For example, at the senior high level (10-12), academic development and physical fitness show up high with indicators of 5, while vocational development and social behavior are low with

TABLE 1

HYPOTHETICAL MATRIX OF PERFORMANCE INDICATORS FOR FOUR LEVELS OF A SCHOOL SYSTEM USING FIVE OUTPUT MEASURES

Output Level	Self Under-Standing Self Acceptance	Academic Develop-ment	Social Behavior	Vocational Development	Physical Fitness
10 - 12	3	5	2	2	5
7 - 9	4	5	2	4	4
4 - 6	2	3	3	2	5
1 - 3	1	5	4	2	5

indicators of 2. Overall, physical fitness and academic development seem to be the strong points of the system. In addition, the matrix seems to indicate that the system is more effective at some levels than it is at other levels. For example, the junior high level (7-9) seems to be doing a better job in promoting student self concepts, than is the primary level.

In summary, the Dyer Model views output as a function of input and hard-to-change surrounding conditions. Performance indices are calculated for each school or system along a number of output dimensions. Schools with higher performance indices are judged more effective than schools with lower ones for specified values of the predictors. Once effective schools have been identified, the easy-to-change surrounding conditions and the process variables are investigated in order to provide clues as to why these schools are more effective than the others.

Production Process Model

The Production Function originates in the economic literature and was first applied to school effects studies by Burkhead, Fox, and Holland (1967) in their study of input and output relationships in large city school systems. Since then the Production Function has been applied to school studies by Hanushek (1970, 1972), Hanushek & Kain (1972) and Levin (1970).

Basically, the model can be represented by the following:

$$A_{it} = f\ [\ F_{i(t)}, S_{i(t)}, P_{i(t)}, C_{i(t)}, I_i\ ]$$

where, A_{it} represents achievement of student i at time t;

$F_{i(t)}$ represents the individual and family background variables averaged over the time interval of the study;

$S_{i(t)}$ represents the school characteristics averaged over the time interval of the study;

$P_{i(t)}$ represents the peer group variables averaged over the time interval of the study;

$C_{i(t)}$ represents the community or external influences averaged over the time interval of the study;

I_i represents the initial student achievement measure or the innate student ability;

f represents the functional relation of all the predictor variables to achievement.

Thus, achievement is viewed as a function of family background, school characteristics, community and external influences, and initial student achievement and/or innate ability.

A clarification of some of the elements of the model is in order at this point. Achievement is the output variable in the above representation. Actually, any output variable consistent with the goals of the schools to be compared could be used in place of achievement.

Individual and family background variables would consist mainly of socio-economic (SES) variables and include such considerations as parents' education, family income, father's occupation, type of goods in the home, location of neighborhood, family size, parents' aspirations and attitudes, and so on.

School characteristics include such variables as teacher characteristics (average age, salary, experience, education, etc.), school resources (audio-visual equipment, library facilities), administrative characteristics (philosophy of principal, guidance services, discipline procedures), and so forth.

Peer group inputs refer to aggregates of the family background measures of the other students in the school, especially their educational and occupational aspirations and expectations. The need for this variable became clear in the Equality of Educational Opportunity Study by Coleman, Hobson, McPartland, Mood, Weinfeld, & York (1966). Hanushek (1972) also found this variable quite important in his study.

Community or external influence variables include the type of neighborhood in which the school is located, the attitude of the community toward education, the tax rate to support education, the amount of community involvement in the schools, and so forth.

Student input refers to ability measures or initial measures of the type of variable being considered as outcome.

The Production Process Model recognizes that education can be viewed as a process in which various variables, both individually and jointly, act to produce outcomes. Users of this model are most interested in identifying the variables that can be manipulated so as to affect certain outcomes. Thus, for policy purposes, family background variables, community variables, student inputs, and, to a lesser extent, peer group variables are not as interesting as school variables, since they are not amenable to direct manipulation. It is the class of school variables and how they relate to the educational process that most interest the policy makers.

Thus, the Production Process Model is very closely related to the Dyer Model. School characteristics variables would include what Dyer calls process variables and also some of the surrounding condition variables, both easy and hard to change. Both models recommend studying the process variables, the easy-to-change condition variables, and the manipulatable school characteristic variables as the key to increasing effectiveness. The models differ in their determination of outcome. The Production Process Model uses all variables to determine outcome, while the Dyer Model uses only input and hard-to-change conditions.

The researcher is encouraged to use one or both of these models as a general strategy in approaching a school effectiveness study. Both seem to provide logical guides as to what classes of variables the researcher might consider, and how each of these can be used effectively in the determination of effective schools. The remainder of this paper deals with the statistical tools and other technical considerations which are necessary to carry out such studies. Once the researcher has possession of these tools, both of these general models should prove useful in providing guidelines for the completion of a school effectiveness study.

THE STATISTICAL MODELS

Six models are proposed as plausible ways to estimate school effectiveness indices. Each model will be briefly described along with an explanation as to how it would be used to determine effectiveness. After all of the models have been presented, each will be discussed separately. Throughout this section, in the interest of uniformity and simplicity, the following notation will be employed:

Y - represents the dependent variable or the measure of the particular outcome under consideration. When multiple outcomes are considered, a vector of dependent variables is appropriate and can be denoted by $\underline{Y}$.

Y - represents the family of input variables or initial status measures on the student for any given outcome. When more than one input is used for a single outcome, X can represent a collection of X_i's, $i = 1,n$, where n is the number of inputs for each outcome. When a vector of single inputs is used with a vector of outcomes, the notation used is $\underline{X}$. When a vector of inputs each consisting of more than one input is needed, the notation is $\underline{X_i}$.

W - represents the collection of the families of all the other variables identified by the general model known to be related to the outcome. Thus, W includes measures of family background variables, school variables, peer group variables and community variables. Ordinarily, several particular measures of these variables will be included in the equations used, therefore W usually represents several such W_i. When a vector of outcomes $\underline{Y}$ is used, W is represented by the corresponding vectors $\underline{W}$ or $\underline{W_i}$, whichever is appropriate.

The relationships that follow will be written in terms of Y, X, and W without subscripts or vector notation. However, the reader should keep in mind that X and W represent collections of variables, and ordinarily more than one variable from each collection will be included in the relationships. Likewise, it is possible for multiple dependent measures (outcomes) to be used. In this case, Y, X, and W all are vectors.

Finally, some school and community variables are often defined so as to be constant for all students in a given school. For example, if the tax rate for the community in which the school was located were used as a predictor, the same value would be used for all the students in the school. Such constant predictors cannot be used in those models (1, 2, 3, and 5) which use individual student scores instead of school means. If it is not possible to redefine these predictors so as to allow some variation, they must be deleted from these models.

Model 1: Analysis of Covariance (ANCOVA)

For each school, a prediction equation is obtained from the regression of the individual outcome scores Y on the appropriate covariates X and W, under the constraint that the least squares estimates for the coefficients of the covariates, b and c, are the same for each school:

$$Y' = a + b\,X + c\,W$$

where, Y' is the predicted outcome for an individual student,

X and W are measures of the respective covariates for that individual, b and c are the least squares estimates of the coefficients of the covariates (in general, these will be vectors),

a is the least squares estimate of the intercept for the school.

The intercept, a, will most probably be different for each school and can be calculated for each school as follows:

$$a = \overline{Y} - b\,\overline{X} - c\,\overline{W},$$

where, Y, X, and W are the respective means for the particular school on the outcome measure, the input measure, and the school measure.

Because of the assumption that the coefficients of the covariates are the same for each school, the planes (lines, if only one covariate is used) obtained for each school will be parallel. Therefore, the difference between the intercepts for two different schools can be used as an effectiveness index. Schools i and j have different effectiveness indices if the hypothesis:

$$H_o: \; a_i = a_j$$

can be rejected. The significance test is standard (Winer, 1971, p. 772).

Model 2: Within-School Regression (Non-standard ANCOVA)

For each school, a prediction equation is obtained from the regression of the individual outcome scores Y on the appropriate predictors X and W:

$$Y' = a + b\,X + c\,W.$$

Here no assumption is made concerning the coefficients of the predictors being the same for each school. The planes (lines) obtained under this model for each school will not be parallel. Hence this model allows schools to be tested for differential effectiveness at various values of X and W, say X_o and W_o.

An effectiveness index can be defined for each school as follows:

$$E.I. = \overline{Y} - b\ (\overline{X} - X_o) - c\ (\overline{W} - W_o)$$

where, (X_o, W_o) is the point at which the different schools are to be compared for effectiveness. Normally, several such ordered pairs will be of interest to the researcher. For the same reference point, two effectiveness indices are significantly different if the confidence intervals on the two regression lines do not overlap (Draper & Smith, 1966, pp. 22-23).

Model 3: Within-School Regression Corrected for Unreliability of the Predictor Measures (Corrected Non-standard ANCOVA)

This model is the same as Model 2, except that the least squares estimates of the coefficients of X and W are corrected for the unreliability present in these measures. The correction is made by dividing the coefficients by the reliability of the measure of the predictor (McNemar, 1969). The effectiveness index is then defined as follows:

$$E.I. = \overline{Y} - \frac{b}{r_{xx}}(\overline{X} - X_o) - \frac{c}{r_{ww}}(\overline{W} - W_o)$$

where, r_{xx} and r_{ww} represent the reliability of the measures X and W, respectively. Because of the correction for unreliability, no standard test is available to determine the difference between two E.I.

Model 4: Mean Difference Scores (Raw Gain)

For each school, the difference between the mean of the input measure X and the mean of the outcome Y is obtained:

$$E.I. = \bar{Y} - \bar{X}.$$

This is the average raw gain from initial to final status for each school. The test for determining whether two E.I. are significantly different can be found in McNemar (1969, pp. 97-98).

Model 5: Individual Regression Residuals

For the total group, a prediction equation is obtained from the regression of the individual outcome scores Y on X and W:

$$Y' = p + q\,X + r\,W$$

where, Y' is the predicted outcome for an individual student,

X and W are measures of the respective predictors for that individual

q and r are the least squares estimates of the coefficients of the predictors based on the total group (once again, in general, q and r will be vectors),

p is the least square estimate of the intercept of the regression line based on the total group.

The residuals for individuals are obtained by subtracting the observed outcome Y from the predicted outcome Y'. The effectiveness index for each school is then calculated by averaging the residuals for the individuals in that school. Symbolically:

$$\begin{aligned} E.I. &= \frac{1}{k} \sum_{i=1}^{k} [Y_i - (M_y - q\,M_x - r\,M_w + q\,X + r\,W)] \\ &= (\bar{Y} - M_y) - q\,(\bar{X} - M_x) - r\,(\bar{W} - M_w) \end{aligned}$$

where, M_y, M_x, and M_w are the means for the total sample on each of the measures; $\overline{Y}$, $\overline{X}$, and $\overline{W}$ are the means for the particular school on each of the measures; and k is the number of students measured in the particular school. No standard test is available to test the difference between two E.I. The researcher may choose to use an adaptation of the test suggested for Model 6 by Dyer, Linn & Patton (1967, pp. 58-59).

Model 6: School Regression Residuals

For the total group, a prediction equation is obtained from the regression of the mean outcome $\overline{Y}$ for each school on the mean predictors $\overline{X}$ and $\overline{W}$ for each school:

$$\overline{Y}' = p' + q' \overline{X} + r' \overline{W}$$

where, p', q', and r' are the least squares estimates of the coefficients when means are used instead of individual observations. These will generally be different from the coefficients obtained in Model 5. The residuals for schools are obtained by subtracting the observed mean outcome $\overline{Y}$ from the predicted mean outcome $\overline{Y}'$. The residual is used as the effectiveness index. Symbolically,

$$E.I. = (\overline{Y} - M'_y) - q' (\overline{X} - M'_x) - r' (\overline{W} - M'_w)$$

where, M'_y, M'_x, and M'_w are unweighted averages over the schools on the particular measure. No standard test is available to test the difference between two E.I. Dyer, Linn & Patton (1967, pp. 58-59) outlined a procedure for testing this difference.

CRITIQUE OF EACH STATISTICAL MODEL

Model 1: ANCOVA

The analysis of covariance was introduced by Fisher (1932) to handle situations in which intact groups were used, that is, when subjects were not randomly assigned to groups. Fisher required that the treatments be randomly assigned to the groups, however, in order to assure that the relationship between the covariate and treatment levels was no more than chance. A crucial assumption of ANCOVA, the homogeneity of covariate coefficients from group to group (parallel lines, planes, or surfaces), was thus expected to hold in most cases (Evans & Anastasio, 1968). Gross violations of this assumption invalidates the analysis (Elashoff, 1969; Winer, 1971); however, McNemar (1969) claims that probably minor violations are tolerable. Both McNemar (1969) and Winer (1971) recommend that ANCOVA be used with intact groups only when it is possible to randomly assign treatments to groups.

In the past, ANCOVA has been used with intact groups where treatments were not randomly assigned to the groups. Werts & Linn (1969) point out that most school effects studies fall into this category. The groups are the students and the treatments levels are the schools, and schools are not usually randomly assigned to students. In fact, there is some evidence to believe a systematically biased assignment is usual (Michelson, 1970; Spady, 1973). In these situations a strong relationship usually results between the treatment effects and measures of the covariate (Evans & Anastasio, 1968). When this occurs, the assumption of homogeneity of regression is generally untenable. The danger of using ANCOVA in this situation is being recognized more and more in the literature recently. (See, for example Antiqullah, 1964; Campbell & Erlebacher, 1970; Cronbach & Furby, 1970; Elashoff, 1969; Lord, 1967, 1969; McNemar, 1969; Werts & Linn, 1971; and Winer, 1971.)

Note that the problem here concerns the relationship between the covariate and the treatment levels, and not the relationship between the covariate and the outcome. In fact, a high correlation between the covariate and the outcome is usually desirable.

Sprott (1970) attempted to soften this criterion by arguing that the criterion should be that treatment is known not to influence the covariate. The expected value of the correlation between the treatment and the covariate is zero in the population. However, Harris, Bisbee, & Evans (1971) claim that this argument rests on an unconventional random effects ANCOVA model.

On the other hand, Jennings (1972) and O'Connor (1972) recommend the use of ANCOVA over analyses based on change scores or residual gain (Models 4, 5, and 6). O'Connor claims that using change scores or residual gain scores gives the same results as ANCOVA, or results that are more difficult to interpret.

In summary, Model 1 appears to be only of limited use due to the assumption of homogeneity of regression. When dealing with existing groups of students in schools, this assumption is particularly restrictive. Furthermore, the more schools one has or the more covariates that are used, the more untenable is the assumption. Uncritical use of Model 1 is to be avoided, unless the researcher is able to determine from his data that the departures from the assumption of homogeneity of regression are not gross. It appears that a superior alternative is available, and that is Model 2.

Model 2: Non-standard ANCOVA

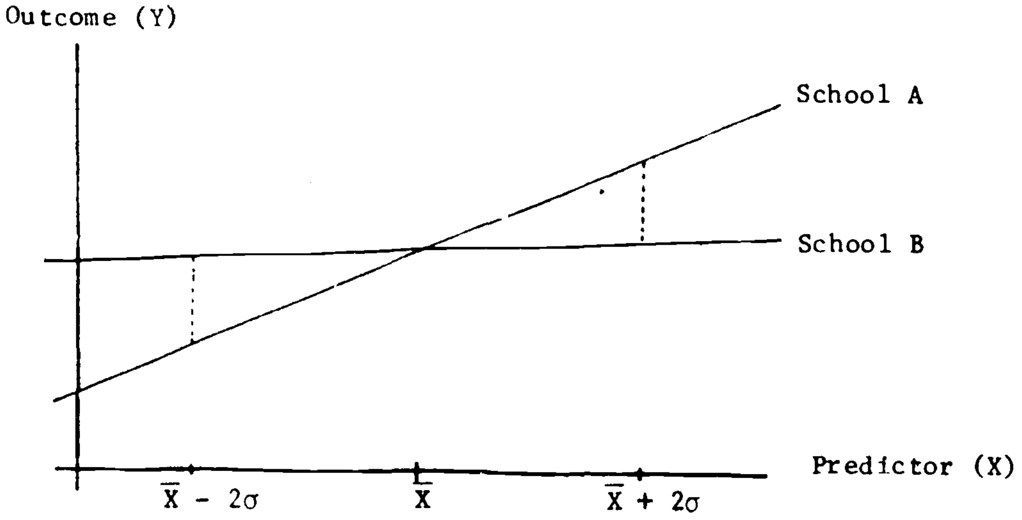
Model 2 is similar to Model 1, except no assumption is made about the parallelism of the planes from school to school; that is, the assumption of homogeneity of regression is not required for this model. In the preceding section it was noted how restrictive this assumption is and how untenable it is in most school effectiveness studies.

Model 2 does not permit the calculation of a single overall effectiveness index for each school. Because the planes representing each school are not necessarily parallel, the intercept is not to be used as an overall effectiveness index. Instead, many effectiveness indices are possible for each school. These are all contingent upon what values of the predictors are inserted into the model. This allows for the possibility that some schools are better for students high on one or more predictors than for those that are low on these predictors. This seems to be consistent with reality and prior research findings (Dyer, Linn & Patton, 1969).

A simple example may help to clarify this point. Suppose that only one predictor is used. Each school is represented by a line, and the coefficient of the predictor is the slope of that line. Dyer, Linn & Patton (1967) note that schools in which students low on the predictor show larger gains than those high on the predictor will be represented by a line with a relatively flat slope. If the gains are larger for students high on the predictor, the slope of the line will be relatively steep. In Figure 1, for example, School A appears to be more effective

FIGURE 1

WITHIN-SCHOOL REGRESSION LINES FOR TWO SCHOOLS



than School B for students high on the predictor ($\bar{X} + 2\sigma$), while School B appears to be more effective than A for students low on the predictor ($\bar{X}$ - 2). It appears from Figure 1, that for some range of scores about the mean on the predictor ($\bar{X}$), both schools are equally effective. Thus, any attempt to produce a single index for School A for all students and one for School B for all students would be misleading.

The use of this model is particularly advantageous when one is interested in comparing schools only at selected points of interest. For example, consider a situation in which initial achievement, SES, teacher experience, and wealth of the community are the predictors. A researcher might be interested in determining the relative effectiveness of schools for students who are one standard deviation below the mean on initial achievement and on SES, and who have a teacher with an experience measure one standard deviation above the mean in a school in a neighborhood of above average wealth. He might expect different results than if he looked at the same students with teachers who have an experience measure one standard deviation below the mean in schools in the same type of neighborhood. Model 2 will allow such comparisons, and for that reason it is particularly appealing.

Dyer, Linn & Patton (1967) contemplated using Mcdel 2 in a pilot study to test some of the technical questions concerned with the Dyer Model. Howevar, probably because of the massive data collection involved, they did not use the model in the study. Herein lies a major problem with this model. A measure for each studenc on each outcome and predictor must be obtained. Therefore, a search of student records is a must, along with the possibility that if certain information is not present in the records, questionnaires to parents, schools, and community officials would have to be employed.

Another problem is the possibility that effective predictors for one school are not effective predictors for another. One alternative here is to use only those predictors found effective for each school. This is not a satisfying alternative, since schools with a different set of predictors cannot be compared (their planes will be in different spaces). A second alternative is to use as predictors for every school any of those found effective for a given school. This enables direct comparison of all schools; however, the number of predictors may be large. This will undoubtedly lead to difficulties in interpretation and sample size. A third alternative is to use only those predictors common to each school. Again, direct comparisons can be made between the schools. However, problems of model specification[1] enter here along with the outside poss'bility (not likely, however) that no common effective predictors can be found. A fourth alternative is to decide _a priori_ what predictors will be used. This has been the approach used in the past, and the one predictor selected has been initial status (cf. Rock, Baird & Linn, 1972; and Marco, 1973). For most initial school effectiveness studies this alternative is probably the most feasible; however, problems with model specification will undoubtedly be encountered.

Another problem involves the selection of the comparison points. In the case of one predictor, such choices as the mean, or one standard deviation above or below the mean may be reaso^able. The problem becomes more complicated when several predictors are used, since an independent decision must be made concerning each one, giving rise to many possible combinations. The researcher should provide some rationale for his choices. If he does not, his results can be challenged simply by noting that his choices were inappropriate.

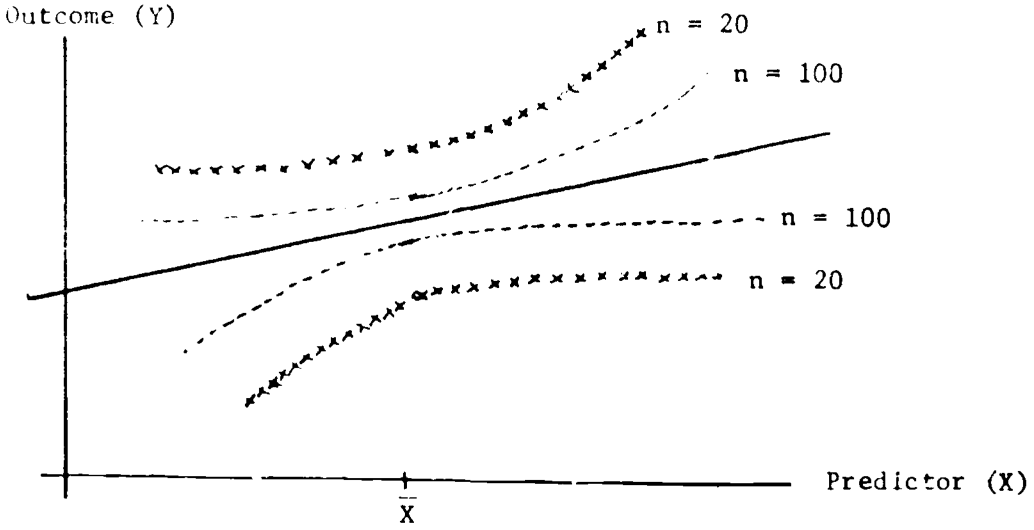
Yet another problem with the model concerns the accuracy of the prediction. The number of predictors, the sample size, and the distance that any component of a particular reference point is away from its respective mean all influence the accuracy of prediction (cf. Draper & Smith, 1966, pp. 22-23). For example, consider the one predictor situation depicted in Figure 2. The dotted lines represent the confidence intervals about the prediction line. Note that the further away the reference point is from the mean, the wider the confidence interval.

[1] Model specification is discussed later in the paper.

Also, the smaller the sample size is, the wider the confidence interval.

FIGURE 2

CONFIDENCE INTERVALS BASED ON SAMPLES OF 20 AND 100 FOR A LINEAR PREDICTION LINE



This situation has some implications for the calculation of effectiveness indices. If two confidence intervals overlap, then it is not possible to state that the two schools have different effectiveness indices. If the samples are small, especially in the situation where several predictors are used, it may not be possible to get confidence intervals which do not overlap. This is especially true as the reference points chosen move farther away from the mean. In these cases, the effectiveness index as defined by Model 2 would not operate efficiently.

Another consequence of this problem occurs when the stability of the indices is investigated. Stability here simply means that if school A shows up more effective than school B with one sample, then, given that all circumstances are the same, school A should show up more effective than school B if another sample is used. Stability has been approached in two different ways in the literature. Dyer, Linn & Patton (1969) and Marco (1973) have used the method of randomly dividing the sample in half and computing indices based on each half. Forsyth (1973) used samples based on two consecutive years to estimate stability. Marco was the only researcher to estimate the stability of indices using Model 2. He found that the estimates were rather unstable with high and low ability students (measured by one standard deviation above and below the mean on input, respectively). This instability is most probably due to the width of the confidence intervals at those particular points, especially since some of his samples contained as few as 17 students.

In conclusion, Model 2 is intuitively pleasing since it allows indices to be computed for selected values of predictors. Also, it provides for differential comparison of schools at these points. This seems to be more in keeping with reality. Furthermore, no restrictive assumptions are placed on the model. The problems with the model center around data collection, specification of predictors, choice of points of comparison, sample size, and stability of the estimates. If the researcher is able to solve most of the problems satisfactorily, it appears that Model 2 can be very useful in determining the relative effectiveness of schools.

Model 3: Corrected Non-standard ANCOVA

In classical linear prediction theory, the predictor variables are considered fixed and measured without error (Winer, 1971). This being the case, it seems that correction for unreliability in measures assumed to have no measurement error is contradictory. Realistically, however, the researcher knows that the predictors are not measured without error. He may even be able to estimate how unreliable his measures are. Thus, the researcher is faced with a decision: should the coefficients be corrected or not?

It is well known that unreliability in any independent variable will bias the weights of all variables toward zero (Cain & Watts, 1968, 1969; Hanushek, 1970; Werts & Linn, 1970; Werts & Watley, 1969). For this reason, Bereiter (1963), Linn, Werts, & Tucker (1971), and Werts & Watley (1969) suggest that the weights be corrected using the usual formula for attenuation (McNemar, 1969). O'Connor (1972) argues that the weights should be corrected only if the object is to interpret the contribution of the variables. However, if the interest is merely to predict the criterion, then the weights should not be corrected. This latter interpretation seems to be keeping with the strict classical view of linear prediction.

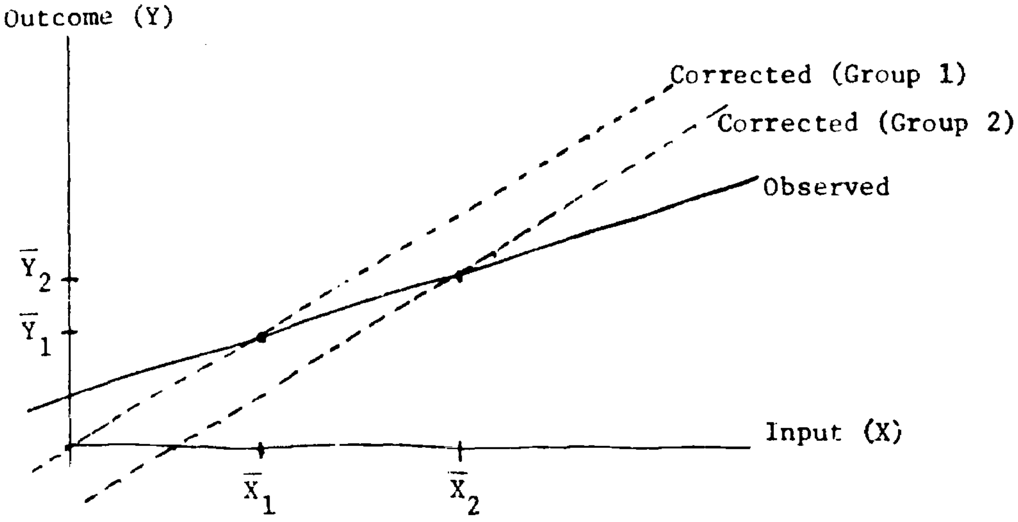
When the researcher is interested in comparing groups not formed at random, Cronbach & Furby (1970) argue for the regression of true outcome status on true predictor status. Essentially, Model 3 does that.

Marco (1973) included Model 3 among several others in computing effectiveness indices. For his data, he found that the results obtained with Model 3 did not deviate appreciably from those obtained with Model 2. However, the reliability of the one predictor that he used in the study was .97. Thus, the correction for unreliability was negligible.

In his report, Marco did present an interesting rationale for correcting the weights. Consider two groups which have the same observed slopes and intercepts (thus, the same prediction line), but different input and outcome means as in Figure 3. When the slopes, and therefore the intercepts, are corrected for unreliability in the input measures, the slope of each line will increase, and the expected value for the group with the lower mean will be higher for any reference point. Without correcting, the two schools would be judged equally effective. When the correction is made, if it is substantial enough, the school with the lower input mean will be judged more effective.

FIGURE 3

COMPARISON OF "TRUE" AND OBSERVED PREDICTION LINES FOR TWO GROUPS WITH DIFFERENT INPUT AND OUTCOME MEANS



Thus, the researcher is faced with a dilemma. If he decides not to correct and assumes error-free meansures, he is in agreement with classical theory, but probably not with reality. If he decides to correct, any test of significance made on the corrected values has no foundation in the linear prediction theory and may be precarious. Probably the best approach is the conservative one of assuming error-free measures and proceeding as though these were present. In this case, Model 3 gives way to Model 2. However, if the researcher decides to use Model 3, he should admit in his report that the statistics used to test for differences have no foundation in linear prediction theory.

Model 4: Mean Difference Scores

At first glance, Model 4 is rather appealing for assessing school effectiveness, since it makes use of a direct measure of change from initial to final status. This appears to be exactly what the researcher desires. Effective schools obviously produce more change than ineffective ones, therefore the direct assessment of change should be an ideal way to determine effectiveness.

Despite their intuitive appeal, measures of raw gain have some difficulties associated with them. Thorndike (1924) was the first to indicate that change scores are generally correlated with initial status. Most of the time the correlation is negative. Thus, schools with low mean scores on the input will have an advantage under this model, since they are likely to gain more. Of course, if a positive correlation happens to exist, schools with high mean scores on the input will have the advantage. O'Connor (1972) notes that such positive correlations are rarely found. The warning about use of measures of raw change appears frequently in the literature (see, for example, Bereiter, 1963; Cronbach & Furby, 1970; Glass, 1968; Jennings, 1972; Lord, 1963; O'Connor, 1972; Rosa, 1972; Traub, 1967; Webster & Bereiter, 1963; and Werts & Linn, 1970).

Cronbach & Furby (1970) note that raw change scores are systematically related to any random error of measurement. They argue that change scores are rarely useful, and advise strongly against their use. Lord (1963) notes that the bias in change scores is not likely to be large, unless the number per group is small. However, he concludes that it is better to avoid their use.

Marco (1973) implied that sometimes the bias in change scores can lead to an unbiased measure of effectiveness. This would occur when this bias counterbalanced bias from other sources (see, also, Campbell & Erlebacher, 1971). However, he offers no criteria as to when this happens and how to detect it. Thus, his argument appears rash.

The use of mean difference scores as a measure of effectiveness has not been widely used in school studies. Dyer, et al. (1967) considered using difference scores rather than residuals in their pilot study. However, the published results of the pilot study (Dyer, et al., 1969) indicated that residual scores had been used.

Marco (1972) used this model in his comparison of effectiveness indices. He found a negative correlation (-.10) between the mean difference score and the initial status mean. He also found that the correlations between the indices determined by this model and those determined by Model 6 was .996. Marco also investigated the stability of the indices as determined by random halves and found satisfactory results. However, his use of students somewhat below average on initial status and the short time interval of six months between initial and final measures may have biased the results.

In summary, the use of change scores has been advised against constantly in the literature. Bias is present, and how the bias will affect the results in unknown. Furthermore, a fundamental assumption in using this model is that in the absence of treatment (i.e., schools) absolute gain would be the same for all students (Campbell & Erlebacher, 1970). In school effects studies this assumption is hardly met due to the other influences that are known to affect output. Hence, use of this model is not recommended despite its intuitive appeal. If the researcher decides to use it, the results should be interpreted with caution.

Models 5 and 6: The Residual Models

The Individual Regression Residual Model (Model 5) and the School Regression Residual Model (Model 6) will be discussed jointly because of their similarities. The two models are basically the same, except that in the former, individual scores are used and the obtained residuals averaged for each school to obtain an effectiveness index, while in the latter school means are used and the obtained residuals are the effectiveness indices. O'Connor (1972) shows by a simple proof that the results obtained from each method will not be identical, hence the need for both models.

The use of residual models in determining effectiveness indices seems appropriate, since the use of a residual score is primarily a way of singling out individuals who changed more or less than expected. This is exactly the intent when one searches for a measure of effectiveness. The assumption is that more effective schools produce larger amounts of change than expected.

Residual models have been frequently used in the past to determine the relative effectiveness of schools. Dyer, et al., (1969) used both Models 5 and 6 in a pilot study. Using a separate analysis for each of the six dependent measures, they found correlations between the deviations obtained by each model to range from .83 to .98, with a median correlation of .93. On the basis of this they concluded that the methods were basically interchangeable.

Dyer, et al. also studied the stability of the estimated indices by using random halves of each school system considered. They found the residuals from Model 5 to be slightly more stable than those in Model 6. The median correlation in Model 5 was .78; in Model 6, it was .72. Due to the slight difference in results from the two models, the authors recommended the use of Model 6 because of the relative ease of obtaining the necessary measures.

Marco (1973) used both models in his investigation of several effectiveness indices. Using reading scores as the dependent measure, he found the correlation between the two methods to be .96. Marco used ANOVA procedures (Winer, 1971) to estimate the reliability of the indices obtained from each model. The reliability estimate for Model 5 was .85 and for Model 6, .83.

O'Connor (1972) recommends that school means be used to compute residuals for comparing schools. He also notes that it is preferable that all groups have the same size, otherwise the residuals may vary greatly in their variance and reliability. From the results of the above studies, it appears that the application of Models 5 and 6 lead to basically the same results. Model 6 is more practical because of the ease of data collection. Forsyth (1973) and Burke (1972) have used Model 6 in their studies.

The use of residual models has not found general favor in the literature. Jennings (1972) claims that there is no good reason for doing residual gain analysis in place of ANCOVA. Werts & Linn (1971b) counter that it is preferable to adopt a regression approach rather than use ANCOVA with its assumptions violated. However, Jennings notes that it is easy to construct data in which the slopes for the two groups are different such that the results of the residual gain analysis are flatly contradicted by the data. Richards (1966) claims that residual scores are notoriously unreliable and subject to errors of various sorts. However, he does not specify what these errors are. One source of error arises from the unreliability of the observed score. Some unreliability is then introduced into the predicted scores, since they are determined by a line fitting unreliable scores. The unreliability is thus compounded in calculating the residual, the difference between two scores with a degree of unreliability in each of them.

Michelson (1970) even goes so far as to attack the use of the linear model. He claims that what is needed is a method of predicting what an increment in an independent variable will do to outcome, given that the other independent variables are constant. He claims that linear models should not pretend to do this since they perform an averaging function.

In summary, residuals have frequently been used to analyze gain situations. In studies where both individual residuals and group residuals have been used, there has been a high correlation between the results of the two methods. This suggests that the use of school means as input is preferable, since means are more readily available than individual scores, and they are more reliable.

In conclusion, it does not seem possible at the present time to single out any of the models as being best in defining school effectiveness indices. Each of the models appears to have its advantages and disadvantages as noted. Some would argue that it is impossible, from an examination of statistics alone, to state what method should be used to analyze a set of data (Hanushek, 1972; Hanushek & Kain, 1972; Linn, Werts & Tucker, 1971; and Werts & Linn, 1970). Thus, the researcher needs to carefully analyze the situation he is considering in order to select the model he feels to be most appropriate.

Some of the models do seem to have restrictive limitations which greatly impair their use. The researcher who intends to use any of these models should be aware of the limitations and their co sequences.

Finally, none of the models presented has ever been validated in a comparison of schools of known quality. Marco (1973) notes that this needs to be done in order to see if any of the models are capable of detecting real differences in effectiveness.

The next section of the paper deals with sundry technical considerations which are necessary for conducting a school effectiveness study.

OTHER TECHNICAL CONSIDERATIONS

In the introduction, a distinction was made between "school effects" and the "determination of effective schools". The former indicated those aspects of the school which had an impact on student outcomes, while the latter involved the ranking of schools by means of some effectiveness indices. Even though exploration of the latter is the major intent of this paper, the two concepts are inseparable in school effectiveness studies. Effective predictors must be identified, so that they can be included in models used to determine the effectiveness indices. It is appropriate at this point in the paper to investigate how problems in identifying these predictors relate to the determination of which school are more effective than others.

Multiple Linear Regression

The most widely used technique to identify effective predictors has been multiple linear regression (Burkhead, et al., 1967; Coleman, et al., 1966; Hanushek, 1972). Typically, the approach has been that after the independent variables have been entered into the regression system, the regression coefficients are tested for significance. If the coefficients are significant, it is concluded that the variable has a significant effect on outcome. Both standardized and unstandardized regression coefficients have been used for this purpose. The use of standardized weights enables the regression coefficients to be directly compared (Werts & Watley, 1969); however, the stability of such weights depends directly upon the variance of the variables for which they are coefficients. Unstandardized weights enable the researcher to determine what effect a unit change in some predictor will have on the dependent measure. Linn, Werts & Tucker (1971) support the use of unstandardized weights even though they admit that these do not, in general, lead to es mates of the relative importance of the effects. Further support for the use of unstandardized weights comes from Blalock (1963), Linn & Werts (1969), McNemar (1969), Tukey (1954), and Yap (1973).

Multicollinearity

A persistent problem when dealing with multiple predictors in a linear regression system emerges as the degree of dependency among the predictors increases (see, Althauser, 1971; Blalock, 1963; Bowles & Levin, 1968a, 1968b; Darlington, 1968; Gordon, 1968; Farrar & Glauber, 1967; and O'Connor, 1972). Economists and sociologists have named this condition multicollinearity. Basically, as the correlation among the independent variables in the equation increases, the standard error of the regression coefficients becomes large and estimates of these are unstable. Blalock (1963) and Gordon (1968) note that the problem exists even when the interdependence among the predictors is low. Farrar & Glauber (1968) thus prefer to speak of the severity of multicollinearity, rather than its existence.

Multicollinearity is a condition arising from the collective impact of the predictors on each other. Specifically, as the interdependence among the variables X increases, the determinant of X^tX approaches 0. Bowles & Levin (1968b) note that this determinant is strictly an ordinal measure of the degree of multicollinearity. The gradient of the determinant as it varies from 1 to 0 is unexplored. For this reason, the seriousness of the problem is difficult to determine from the sizes of the zero order correlations among the variables.

Hanushek (1972) claims that multicollinearity is a problem only if the point estimates of the parameters are to be used. For example, this would occur if the researcher intended to use the estimates as the means for determining which variables in the system should be retained. A serious degree of multicollinearity may result in an improper interpretation of the contribution of such variables. This would not be a problem, however, if the system were to be used for prediction only. Thus, multicollinearity does not pose a threat to the models used to determine effectiveness indices.

The existence of multicollinearity has implications for the methods used to identify what the relevant predictors are. If variables are simply added to the prediction system until the increase in the multiple R is less than some predetermined value, a problem may exist. Walberg (1971) notes that if the predictor variables are correlated, the effects of these variables are confounded, and those entering late in a series of successive tests are less likely to be significant. More precisely, as redundant predictors are entered into a system their common predictive value gets averaged, in a weighted manner, over all of their regression coefficients. Newton & Spurrell (1967) caution against the use of computer programs without investigating exactly how the variables are selected, since the results can differ considerably depending upon the order of selection. Even though stepwise regression procedures have been defended by Darlington (1968) and Draper & Smith (1966), the proper approach is to try every possible combination of orders. However, if the number of predictors is large, this method mav be wasteful of researcher and computer time, and thus can be verv expensive. If the researcher chooses to use a stepwise regression program, proper checks on the different combinations should be employed.

Partitioning of Variance

Another approach for determining the contribution of the variables is the use of the proportion of total variance accounted for by each variable. Walberg (1971) argues that examination of this may often be more useful and valid than determining the significance of the regression coefficients, since in the latter there is neither random sampling from known populations nor random assignment of response units to treatments. Thus, the probability values for the parameters may be meaningless and statistical inference may be unwarranted. Ward (1969) warns that sometimes nonsense occurs in partitioning the variance. He cites Werts (1968), where some negative components are present. Bowles & Levin (1968a) indicate that, in the presence of multicollinearity, proportion of variance explained by each variable is misleading as an index of importance.

Two alternatives have been offered to the partitioning-of-variance method. The first is commonality analysis (Coleman, 1970; Mayeske, 1970; Mayeske, et al., 1969; Newton & Spurrell, 1967; and Tatsuoka, 1973). Commonality analysis partitions the accounted-for variance into a part uniquely associated with one subset of predictors, a part uniquely associated with the complementary subset of predictors, and a part attributable to either of the two subsets. The last is called the commonality. The technique is an heuristic device for exploring how a large set of variables may be meaningfully partitioned into several subsets. If the commonalities are large relative to the unique parts, the partitioning is not an useful one. It is a signal that the "right" partitioning has not been made, or that good indicators of the factors are not present, or both.

The second alternative to partitioning the variance is factor analysis. Creager (1971) prefers this to commonality analysis since the former is an orthogonal method of grouping the variables while the latter is nonorthogonal. Mood (1971) recommends factor analysis since he claims that individual regression coefficients will seldom give much help in identifying relevant variables. However, Mood notes that in the present state of understanding, factors would have to be selected mostly by intuition. Stephenson & Beard (1971) employed principal axis factor analysis in their study of the school, social, and economic environment in Florida. Dyer (1972b) recommends factor analysis as a variable reduction technique when many variables are available. If the researcher is satisfied with the results of the factor analysis, the factor scores could be used as measures of the predictors in whatever statistical model is selected to determine effectiveness indices.

Partial and Part Correlation

Partial and part correlations have also been employed to determine school effects. In particular, the partial correlation between the school variable and the outcome variable with input controlled has been used. Also, the part correlation between the school variable and the outcome variable with the input partialed out of the school variable or out of the outcome variable has been used. It has been noted often in the literature that partial regression coefficients are superior to either of the above because controls for input may be introduced without underestimating the magnitude of the true effects (see, for example, Astin, 1963; Blalock, 1963; Richards, 1966; Tukey, 1954; and Werts & Watley, 1968, 1969).

Werts & Linn (1969) showed how partial correlations, part correlations, and standardized partial regression weights are related to each other. Furthermore, they demonstrated that the optimal method to use in any study is a function of the hypothesis one wishes to support and the pattern of obtained correlations. In other words, a method is available to support the researcher's biases. This fact alone should prompt a very careful review of the techniques used in any school effects study before the findings can be interpreted correctly.

In summary, the determination of which predictors are important involves many methodological problems.

Model Specification

Once the researcher has decided on the predictors and the model he wishes to use to determine the effectiveness indices, two major sources of error loom as threats to his study. One source of error deals specifically with the choice of the model and is called specification error. The second source of error deals with the measures of the variables and will be discussed later.

There are two forms of specification error. One arises from an inappropriate choice of model to represent the reality of the situation, and the other arises from an improper choice of predictors. Both types will be discussed briefly.

With the exception of Model 4, each of the models discussed is a special case of the general linear model (Cohen, 1968; Winer, 1971). Thus, if any of these are used it is assumed that the proper relationship between the predictors which have been identified as important can be adequately expressed in terms of a least squares additive model. This is almost certainly not an accurate representation of reality. However, under the circumstances, such models are probably the best available at the present time. These models are used because they are familiar and have a strong statistical foundation. However, they are to be used with care, since they are almost certainly inadequate representations of the way the variables act, separately and jointly, to affect the dependent variable.

Given that the general linear model is probably the best available at the present time, which form of the model should be used to specify the relationships? Should only linear terms be used, or will the introduction of product, quadratic, or higher order terms into the model lead to better specification? Before using any one of the models, the researcher ought to obtain a scatterplot of the variables. Usually a model with linear terms provides a good fit unless the relationship is obviously curvilinear. In the event of more than one predictor the problem is more complicated, since more than two dimensions are involved. Walberg (1971) recommends checking linear terms before looking at product and polynomial terms. The gain that may be made in prediction using these terms may not be worth the extra work involved. Richards (1966) and Hanushek (1972) argue for the use of interaction terms, since this is more consistent with their _a priori_ views about the educational process. Hanushek used interaction terms, and he concluded that the statistical properties of the model seemed better than when these terms were not included. His criterion for "betterness" was that the parameter estimates had higher t-values associated with them, thus reflecting greater precision. However, this criterion as a measure of greater precision does not appear to have any statistical foundation. Tuckman (1971) also used these terms, but his results were difficult to interpret due to the lack of a clear pattern.

The second type of specification error is due to improper selection of predictors. Failure to include predictors which influence the dependent variable and which are uncorrelated with the other predictors in the system will generally result in poorer prediction. In addition, failure to include predictors which are correlated with the other predictors in the system will generally result in different estimates for all of the variables studied. The extent of the seriousness of this type of error is not discussed in the literature and probably is unknown. Hanushek (1972) does mention that misspecification is more serious if initial achievement is not included as one of the predictors.

The key to minimizing specification error of this type seems to be to select predictors on the basis of sound theory and prior research (Dyer, et al., 1967; Gordon, 1968; Tatsuoka, 1973; Werts & Watley, 1969). Cain & Watts (1970) claim that the most serious gap in education today is inadequate theory. Hanushek & Kain (1972) encourage radical experimentation in an attempt to uncover proper, and possible hitherto unknown, predictors.

Intimately related to this type of specification error is the problem of imprecise representation of a theoretically justified variable. A measurable representation of a theoretical variable will be called a proxy variable. For example, theory would probably demand that teacher quality be included as a predictor of student achievement. The question then arises as to how teacher quality should be represented. Some proxies of teacher quality that have been used are verbal scores (Coleman, et al., 1966), recency of latest educational experience (Hanushek, 1972), and years of experience (Burkhead, et al., 1967). Are any of these, singly or jointly, adequate representations of teacher quality? Exactly what constitutes teacher quality if largely unknown, and to think that it can be captured by one measure or by a number of measures at our stage of understanding is probably misleading. The same remarks apply to any number of other variables of interest in school effectiveness studies. Hence, specification error is compounded by poor proxies.

In summary, errors due to model specification are almost certainly present in school effectiveness studies. These errors operate in unknown ways and with unknown seriousness. The researcher should be aware of this and attempt to reduce both types of specification error by carefully selecting predictors and proxies according to sound theory and prior research. In addition, he should attempt to choose a statistical model which best captures the reality of the situation as he perceives it. Cain & Watts (1970) warn that the role of a variable in affecting outcome is meaningful and interpretable only in the context of a carefully specified an theoretically justified model.

Measurement Error

The second major source of error in school effects studies is measurement error. Herriott & Muse (1973) note that little attention has been paid to this problem in the studies of educational effects. Dyer (1972b) speaks of the importance of appropriately measuring the variable to be used in the study. To the extent that measurement error is present in the predictors, problems with the assumption of error-free measurement of the classical linear prediction model are encountered. Errors in the dependent measure, even though tolerated by the theory, will result in poorer prediction, and hence work against proper designation of effectiveness indices.

Various sources of measurement error are present in school studies. The major source is the unreliability of the proxy measures used for the independent and dependent variables. Not only are some proxies poor representations of the theoretically justified variables, but many times the measures used to specify the proxies are inadequate. For example, teacher quality may be represented by the proxy teacher experience, which is measured in years of teaching. Is "years of teaching" really a good measure of experience? Dyer (1972b) notes some practical considerations which contribute to error: inconsistencies in the data supplied by the schools, confused record keeping, and the tendency to fill up information gaps with impressionistic or fictitious data.

In summary, measurement error and specification error are closely related. Measurement error existing in the predictors of the model cause assumptions to be violated. Error existing in the dependent variable affects the prediction of effectiveness indices. In general, existence of measurement error almost certainly works to bias the determination of the effectiveness indices.

The Predictors

The researcher must decide which variables to include as predictors in his model. As mentioned previously, the two criteria which should be employed in this choice are theory and previous research. By theory is meant the researcher's conception of how certain variables ought to interact to produce change in outcome. By previous research is meant what variables have been found by others to relate significantly to the outcome under consideration. This section of the paper will deal briefly with the findings of previous research.

Two extensive reviews of school effectiveness studies are available. Guthrie (1970) reviewed 19 school effectiveness studies. He found that in all of these studies, SES seemed to be strongly related to achievement. When SES is controlled, other variables emerge as relating significantly to achievement. It is clear from this review that these other important predictors vary from situation to situation. This variation might be due to the type of school variable investiaged, the type of outcome considered, the sample used, the model which was employed, or the type of external controls used. When searching for important predictors, each of these should be carefully considered. Guthrie includes a summary chart listing the authors, a description of the sample, the outcome measures, and the school variables which emerged as significantly affecting achievement.

Spady (1973) reviewed 12 studies concerned with the impact of school resources on outcomes. Nine dealt directly with financial expenditures, in addition to other variables, and presented somewhat inconsistent results. The remaining three raised questions about the importance of teacher experience and formal training of teachers as predictors of student achievement. Spady concluded that teacher experience must be regarded as an inadequately studied variable whose effect on achievement remains obscure.

Any researcher contemplating a school effectiveness study should begin with these two reviews. In addition to these, suggestions for predictors can be found in Burkhead, et al. (1967), Coleman, et al. (1966), Dyer, et al. (1967), Hanushek (1972), Metfessel & Michael (1967), Stephenson & Beard (1971), Tuckman (1971), and U. S. Office of Education (1971).

The researcher should not blindly accept the findings of such studies, however, when searching for important predictors in his particular situation. These should merely act as guidelines as to what variables may be important. Some problems may be present in these studies which cloud the importance of one or several predictors. Guthrie (1970) pointed out in his review that most of the studies considered did not take into account the student's entering capabilities nor the type of experiences he participated in outside of school. When these are entered into the model, different variables may emerge as important. In fact, Spady (1973) hints that the omission of such variables as intelligence and motivation may inflate the estimated impact of SES on achievement.

A second problem in past studies, as noted by Levin (1970), is that there has been no attempt to specify in a systematic way the particular formulation of how schools affect achievement. The approach has been rather haphazard. When reviewing a study, the researcher should look for some evidence of an underlying theory.

A third problem is the probable confounding of many of the predictors. A good example may be the confounding of SES variables with school variables. Since the study by Coleman, et al. (1966), there has been a concerted effort to show that some school variables are indeed important predictors of achievement. However, when such variables are entered into a model with SES variables, the latter consistently emerge as explaining most of the variation in school achievement or similar outcome measures. The Coleman study has been severely criticized for inadequate measure, inappropriate proxies, and inappropriate statistical techniques (Bowles & Levin, 1968a; Cain & Watts, 1968; Campbell & Erlebacher, 1970; Guthrie, 1970; Michelson, 1970; and Spady, 1973). The criticisms are no doubt valid; however, other reasons may exist to explain the results found by Coleman. Worthington & Grant (1971) argue that the economic and social factors of the area in which the school is located may be reflected in the curriculum, grading standards, and general intellectual atmosphere of the school. Michelson (1970) and Spady (1973) note that a bias exists against finding a significant effect of school resources on outcome which is created by a preselection of neighborhoods and schools by certain groups of people, and the preselection and grouping of students into different ability groups, tracks, and even schools on the basis of their previous achievement and SES. All of these factors contribute to the confounding of the predictors.

Another consideration which should be made by the researcher is that possibly some variables have never even been considered as predictors of achievement or other outcomes. Dyer (1972b) mentions in passing that nutritional and neurological facts that affect growth have never been considered. There may be others.

Thus, the researcher should scan the literature for suggestions and insights. The predictcrs which emerge from these considerations should be tested against the researcher's previously established theory, and either rejected or confirmed. Finally, the variables should be entered into the model chosen by the researcher, and some decision made on their appropriateness.

Input and Output

Since schools are being compared in terms of effectiveness, this implies the use of some criteria probably based on the goals or objectives which the schools hold as important. Thus, the specification of outcome should be along a dimension commensurate with the goals against which the schools are being compared. Levin (1970) argues that if only a single outcome is used, insights are gained only along one dimension and comparisons can be made only along that dimension.

The researcher must be ready to accept that even when a goal can be specified which is common to a group of schools, this goal will probably not receive the same stress in each school. Levin notes that in these cases the relation between any single output and school resources will be underestimated.

Another problem facing the researcher is the choice of an appropriate outcome variable which will represent the goal common to a group of schools. For example, suppose such a goal involves increasing the student's ability to read. Would a vocabulary test be an appropriate measure of this outcome, or should a reading comprehension test be used?

Once that decision is made, it may turn out that the outcome variable chosen may have nothing to do with the determination of what schools are most effective. For example, Hanushek (1972) notes that the use of verbal scores as an outcome measure may be more closely linked to home environment than to the school environment, and thus harder for schools to affect than some other outcome measure.

The proper approach to the identification of efficacious outcome is, once again, through theory and experimentation. However, usually the issue will be decided by convenience; that is, by what measures are possible and practical in a given situation. For example, practically every school used some kind of standardized testing on a regular basis. Use of these measures as outcomes would be convenient for these schools. In addition, often these tests involve the testing of basic skills: reading and mathematics. Therefore, many times these are the bases on which schools are ranked in effectiveness. Measures of social development, physical fitness, and attitudes are not as available, and thus are rarely used as measures of school effectiveness.

Once the outcome measures have been decided upon, what inputs should be used? Must the input measures be exactly the same as the outcome measures? For example, can the Iowa Test of Basic Skills be used as input in the 6th grade and the Comprehensive Test of Basic Skills be used as outcome in the 8th grade? The evidence seems to suggest that only tests of similar structure need to be used, for example, two math computation tests or two vocabulary tests. Campbell & Erlebacher (1972) state that no satisfactory analysis is possible when the pretest is not similar in structure to the posttest. Cronbach & Furby (1970) even question whether the same test given on two different occasions is ever measuring the same thing. Bereiter (1963) hints at the possibility of the same kind of test measuring different things when given at different times, although he states that the meaningfulness of change scores does not depend upon a test's measuring the same thing on two different occasions. Thus, the dilemma is a false one.

In practice, both equivalent and nonequivalent forms have been used. For example, Dyer, et al. (1969) used alternate forms of the Iowa Test of Basic Skills for input and output. Burke (1972) used the Metropolitan Achievement Test as input in the 2nd grade and the Comprehensive Test of Basic Skills as output in the 6th grade.

Despite the convenience and availability of standardized tests, use of these instruments as input and outcome measures for school effectiveness model^ is not without hazards. First, these test generally contain items covering a broad range of curricular objectives, and may not be very good measures of any particular set of objectives. They are usually intended for nationwide use and may not meet the specific needs of local schools for a measure of effectiveness. Tailored instruments developed for use in state testing programs are probably more suitable as measures of effectiveness for those schools in that state.

Second, standardized tests are constructed in such a way that it is difficult for schools to show much gain from input to outcome. Thus, schools scoring high on the input will not be able to gain much on the outcome. This hazard is especially pertinent in Model 4. Also, when equivalent tests are used in Model 4, the reliability of the difference scores is low. Furthermore, as noted in the critique of this model, schools low on the input will tend to show larger gains. This difficulty exists to a lesser extent in the other models which use a predicted outcome approach rather than simply considering raw gain from input to outcome.

In summary, outcome variables which are adequate and practical measures of a dimension on which to validly compare schools as to effectiveness need to be identified. For each outcome measure, a corresponding input measure of at least similar structure needs to be used.

Unit of Analysis

Whether to base the effectiveness study on the comparisons of school systems, individual schools, or individual classrooms depends to a great extent on who wants the study done. A state commissioner of education would probably be most interested in comparing the effectiveness of the different districts in his state. A county superintendent may be interested in comparing the individual schools within his district, while a school principal would probably be more interested in comparing the effectiveness of several different classrooms of the same grade level within his school.

Different problems will be encountered in each situation. As the unit of analysis becomes more encompassing, homogeneity of goals becomes somewhat of a problem. However, an advantage here is that a larger number of pupils are available, and the unit can be compared over a longer period of time. As the unit becomes more restrictive, fewer students are available and the practical constraints of time are present. For example, individual classrooms would probably not remain intact within a school for more than a year, thus prohibiting a longer time interval for the study. This may be crucial since it may take a longer time for the effectiveness of a unit to become manifest. On the other hand, school districts could be compared over a period of several years.

The study reported by Dyer, et al. (1969) utilized school systems as the unit, while the studies reported by Burke (1972), Forsyth (1973), and Marco (1973) utilized individual schools as the unit. No study using individual classrooms as the unit has been found.

Longitudinal and Cross-sectional Data

A longitudinal study is one in which the input and output measures are taken on the same group at two different times. For example, input measures are taken for all 6th-grade students at a certain school, and, two years later, outcome measures are obtained for the 8th-grade students in that school. If the two groups consist of exactly the same students, the data is called matched-longitudinal. If the composition of the two groups is not exactly the same due to additions and/or deletions, the data is called unmatched-longitudinal. Matched-longitudinal data can consist of individual scores or group means, whereas unmatched-longitudinal data must consist of group means.

A cross-sectional study is one in which the input and outcome measures are obtained for different groups, usually simultaneously, or at least within the same school year. For example, measures are obtained for the 6th- and 8th-graders of a certain school during the same year. The 6th-grade measures constitute input, and the 8th-grade measures, the outcome. Group measures are required with this type of data.

The primary advantage of cross-sectional data is the relative ease of obtaining it. Since the measures can be obtained during the same year, a painstaking search of student records or the need to wait several years between measures can be avoided. If the researcher wishes to use cross-sectional data in school effectiveness studies, he must assume that the outcome group, when they were at the same level as the present input group, would perform the same way on the input measures as the present input group performed. Failure to meet this condition would certainly lead to erroneous interpretation of relative effectiveness. Realistically, the researcher probably has no way to check on whether this condition is satisfied, for if he did he would almost certainly use the data in a longitudinal sense.

Despite this limitation, Herriot & Muse (1973) note the presence of this type of data in school effectiveness studies. In particular, Tuckman (1971) used cross-sectional data in an input-output model. Hanushek & Kain (1972) argue that the use of cross-sectional data clearly tends to underestimate the total effects of educational inputs on achievement. However, they do claim that some information can be obtained on the usefulness of certain predictors using this type of data, but they caution that the results must be carefully interpreted. Marco (1973) found a correlation of .79 between effectiveness indices obtained by using longitudinal data and estimates of such obtained from using cross-sectional data. However, no cross-validation was performed on a separate sample of schools.

From purely theoretical considerations, longitudinal data would seem superior to cross-sectional data, since the former allow a direct measure of change while the latter do not. However, the loss in precision resulting from the use of cross-sectional data may well be worth the avoidance of tedious data collection procedures necessary with longitudinal data.

Dyer, et al. (1969) investigated the use of the different types of data in their pilot study. Four samples were used in the study: matched-longitudinal using individual scores (sample 1), matched-longitudinal using group means (sample 2), unmatched-longitudinal (sample 3), and cross-sectional (sample 4). Correlations were obtained between the residuals from the regression surface on each of the six outcome measures employed for each of the samples. The two matched-longitudinal samples (1 and 2) had a median correlation of .93. The median correlations among the other possible combinations ranged from -.07 (1 with 4) to .36 (2 with 3). Thus, the results seem to indicate that for the three-year time interval covered by the study, unmatched-longitudinal or cross-sectional samples cannot be relied upon to produce the same results as matched-longitudinal samples.

In summary, the researcher is advised to employ matched-longitudinal data whenever feasible. Both from theoretical and experimental considerations, this type of data appears superior to alternatives. Unmatched-longitudinal data do provide a direct measure of change over time, but the possibility of a radical change in the composition of the group looms as a threat to interpretation of results. Even though cross-sectional data are the most practical from the standpoints of availability and ease of collection, they would appear in most cases to have severe limitations which preclude their use in school effectiveness studies.

Multivariate and Univariate Analyses

For the most part, multiple dependent measures analyzed jointly have not been used in school effectiveness studies. Even though multiple outcomes have been considered by most studies, separate univariate analyses have been conducted on each dependent variable (see, for example, Burkhead, et al., 1967; Coleman, et al., 1966; Dyer, et al., 1969; and Forsyth, 1973). Several reasons for this trend are offered. First of all, many researchers are not acquainted with multivariate techniques. Multivariate analysis is just beginning to make its appearance in educational research. Tatsuoka (1973) indicates that the first treatise on multivariate analysis addressed specifically to educational researchers was by Cooley & Lohnes (1962). Since then notable contributions have been made by Bock (1973), Cooley & Lohnes (1971), Morrison (1967), Tatsuoka (1971), and Van de Geer (1971).

Secondly, even if multivariate techniques are familiar, many researchers are more comfortable with univariate techniques. Univariate computer programs are readily available and easier to use; interpretation of univariate results is usually easier and the researcher is assured that his readers will more readily understand his attempt to convey univariate results instead of multivariate results.

Thirdly, basic misconceptions about multivariate analysis seem to exist among some researchers. For example, Rock, Baird, & Linn (1972) argue that since the overall multivariate F for a particular problem was not significant, further interpretation based on the univariate F's was not warranted. This interpretation completely disregards the fact that the significance of an overall multivariate F is not dependent upon the significance of each univariate F.

The use of multivariate techniques seems ideal for school effectiveness studies. Certainly, the effectiveness of a school needs to be assessed on more than one dimension, thus the outcomes of such a study are multivariate in nature. School effectiveness can thus be considered globally as it relates to all outcomes, and the use of multiple cutcomes jointly allows the variables to be analyzed in such a way so as to make use of the inherent dependency present among the measures. Multivariate analysis can shed light on just how each variable contributed to the overall effect, precisely because the variables are considered simultaneously (Tatsuoka, 1973). The warning appears frequently in the literature than any variable considered in isolation may affect the criterion differently from the way it will act in the company of other variables (Morrison, 1967; Walberg, 1971).

The researcher is encouraged to consider a multivariate analysis of his data when conducting a school effectiveness study. Procedures exist to adapt the models presented in this paper to handle multiple dependent measures. Any of the basic references provided above could be used. In addition, since univariate analysis is a special case of multivariate analysis, the researcher can conveniently obtain the results of the separate univariate analyses, if he so desires. Finally, a warning is in order. Multivariate procedures should not be used blindly. The researcher should be familiar with the procedures and the problems involved with the interpretation of the results. The uncritical, mechanical use of sophisticated techniques that have become possible through the widespread availability of computers and "canned programs" is to be avoided.

CONCLUSION

The researcher who intends to conduct a school effectiveness study is faced with many methodological considerations. This paper has focused on the major considerations and has provided some guidelines, based on logical thinking and prior research, so that the researcher may be able to make informed decisions when faced with these considerations in his study. Unfortunately, since the problem is so complex, no definitive solution to many of the problems encountered can be offered. What is offered is the experience and findings of other researchers who have tackled similar problems, so that these might prove helpful to the researcher when he encounters these same problems in his study.

In the introduction, a distinction was made between studies designed to identify effective predictors of student development, and those designed to identify more effective schools. The close relationship between these two types of studies was noted. The former was viewed as a prerequisite for the latter in most cases. The methods considered in this paper mostly concern the latter type of study.

The Dyer Model and the Production Process Model were proposed as providing a theoretical basis for school effectiveness studies and an overall strategy for conducting them. Both models help identify important classes of variables, and provide a plan as to how each class might be most useful in school studies.

The six models proposed as methods of calculating effectiveness indices were thoroughly discussed and critiqued. The Within-School Regression Model (2) and the two residual models (5 and 6) were found to be the most appealing from a theoretical viewpoint. Model 2 allows for indices to be computed at different levels of input, and for this reason is particularly appealing. However, the data collection problem that it involves and the instability of the indices found in the one study in which this model was used may render it impractical. There is a need for Model 2 to be applied in a carefully designed study to test its practical applicability. Results from Models 5 and 6 have been rather similar. Since Model 6 employs means instead of individual observations, it is recommended for use in situations where the researcher has reason to believe that the results from both models will be similar. The other three models (1, 3, and 4) have severe limitations which will greatly hamper their use in most studies.

The validity of the six models needs to be studied. Until the present, each model has found favor or disfavor due to theoretical considerations or to the amount of agreement shown when applied to real data. There is a need to apply these models to schools of known quality to see if any or all of them are capable of detecting differential effectiveness among schools. Until this is done, the researcher is advised to employ Models 2, 5, and 6, and compare their results.

Selection of the appropriate predictors should be based on sound theory and prior research. However, past research should be carefully scrutinized by the researcher for possible methodological failings. There is a pressing need for a well developed theory of what variables, both individually and jointly, affect student development. In addition, there is the ever present need for well designed studies to identify effective predictors.

The choice of outcome measures depends upon the determination of the dimension along which the schools are to be compared. If valid comparisons are to be possible, the researcher must be able to identify dimensions consistent with the goals and objectives of the schools to be compared. Once the outcome measures are specified, input variables should be of the same type as the outcomes, but measured at an earlier time. Matched-longitudinal data are recommended over unmatched-longitudinal or cross-sectional data despite the more cumbersome collection involved.

Whether the units to be compared are school systems, individual schools, or individual classrooms depends mainly upon the purpose of the study. The size of the unit does have implications for the size of the sample that can be used and the time interval over which the study may be conducted.

When multiple outcomes are considered, a multivariate analysis of the data is superior to separate univariate analyses. However, the researcher should be familiar with such techniques before he attempts to apply them. Mechanical use of canned computer programs should be avoided.

Finally, two major sources of error in school effectiveness studies arise from improper model specification and inadequate measurement. Once important variables have been identified by theory and prior research, there is a pressing need to develop appropriate representations or proxies for these variables, and then to develop adequate measures of these proxies.

In conclusion, research is needed to identify effective predictors of student development and to determine which models designed to produce effectiveness indices are valid. The identification of effective predictors includes the development of an adequate theory of what does affect student development, adequate representation of the variables so identified, and adequate measures of these representations. Some research has been going on in this area in the past several years, but without much success. Still relatively little is known about what affects student change. There has been no research done on the validity of the models for calculating effectiveness indices. With a concerted effort on the part of educational researchers, sociologists, economists, and other researhers, hopefully some of these problems will be solved in the near future. Schools do make a difference. Let's find out why.

BIBLIOGRAPHY

Aigner, O. J. A comment on problems in making inferences from the Coleman Report. American Sociological Review, 1970, 35, 249-252.

Altnauser, R. Multicollinearity and non-additive regression models. In H. M. Blalock (Ed.), Causal model in the social sciences. Chicago: Aldine-Atherton, 1971.

Astin, A. W. Differential college effects on the motivation of talented students to obtain the Ph.D. Journal of Educational Psychology, 1963, 54, 63-71.

Atiqullah, M. The robustness of the covariance analysis of a one-way classification. Brometrika, 1964, 51, 365-372.

Bereiter, C. Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wis.: University of Wisconsin Press, 1963.

Blalock, M. Correlated independent variables: the problem of multicollinearity. American Journal of Sociology, 1963, 42, 233-237.

Blalock, M. Causal models in the social sciences. Chicago: Aldine-Atherton, 1971.

Bock, R. D. Multivariate statistical methods in behavioral research. New York: McGraw-Hill, 1973.

Bowles, S. and Levin, H. The determinants of scholastic achievement-an appraisal of some recent evidence. Journal of Human Resources, 1968, 3, 3-25.

Bowles, S. and Levin, H. More on multicollinearity and the effectiveness of schools. Journal of Human Resources, 1968, 3, 393-400.

Burke, H. R. A study in public school accountability through the application of multiple regression through selected variables. Unpublished doctoral dissertation, Indiana University, 1972.

Burkhead, J, Fox, T., and Holland, J. Input and Output in large-city High Schools. Syracuse: Syracuse University Press, 1967.

Cain, G. and Watts, H. The Controversy about the Coleman Report: Comment. Journal of Human Resources, 1968, 3, 389-392.

Cain, G. and Watts, H. Problems in making policy inferences from the Coleman Report. American Sociological Review, 1970, 35, 228-242.

Campbell, D. T. and Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In Hellmuth, J. (Ed.), Disadvantaged child: compensatory education a national debate. Vol 3. New York: Brunner/Magel, 1970.

Cohen, J. Multiple regression as a general data analytic system. Psychological Bulletin, 1968, 70, 426-443.

Coleman, J. S. Communications: equality of educational opportunity: reply to Bowles and Levin. Journal of Human Resources, 1968, 3, 237-246.

Coleman, J. S. Reply to Cain and Watts. American Sociological Review, 1970, 35, 242-249.

Coleman, J. S., Campbell, E., Hobson, D., McPartland, J., Alexander, M., Weinfeld, F., and York, R. Equality of educational opportunity. (OE 38001) Washington, D. C.: U. S. Department of Health, Education, and Welfare, 1966.

Cooley, W. W. and Lohnes, P. R. Multivariate procedures for the behavioral sciences. New York: Wiley, 1962.

Cooley, W. W. and Lohnes, P. R. Multivari te data analysis. New York: Wiley, 1971

Creager, J. A. Orthogonal and nonorthogonal methods for partitioning regression variance. American Educational Research Journal, 1971, 8, 671-676.

Cronbach, L. and Furby, L. How we should measure 'change' - or should we? Psychological Bulletin, 1970, 74, 68-80.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Draper, N. and Smith, H. Applied regression analysis. New York: John Wiley, 1966.

Dyer, H. The Pennsylvania Plan. Science Education, 1966, 50, 242-248.

Dyer, H. Toward objective criteria of professional accountability in the schools in New York City. Phi Delta Kappan, 1970, 52, 206-211. (a)

Dyer, H. Can we measure the performance of educational systems? National Association of Secondary Schools Principals Bulletin, 1970, 54, 96-105. (b)

Dyer, H. The measurement of educational opportunity. In Mosteller, F. & Moynihan, D. P. (Eds.), On equality of educational opportunity. New York: Random House, 1972. (a)

Dyer, H. Some thoughts about future studies. In Mosteller, F. & Moynihan (Eds.), On equality of educational opportunity. New York: Random House, 1972. (b)

Dyer, H. School evaluation: a realistic response to accountability. North Central Association Quarterly, 1972, 46, 390-396. (c)

Dyer, H., Linn, R., and Patton, M. Feasibility study of educational performance indicators: final report to New York State Education Department. Princeton, N. J.: Educational Testing Service, 1967.

Dyer, H., Linn, R., and Patton, M. A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 1969, 4, 591-605.

Elashoff, J. Analysis of covariance: a delicate instrument. American Educational Research Journal, 1969, 6, 383-403.

Evans, S. H. and Anastasio, E. J. Misuse of analysis of covariance when treatment effect and covariate are confounded. Psychological Bulletin, 1968, 69, 225-234.

Farrar, D. and Glauber, R. Multicollinearity in regression analysis: the problem revisited. Review of Economics and Statistics, 1967, 49, 92-107.

Firman, W. The quality measurement project in New York State. Science Education, 1966, 50, 259-297.

Fisher, R. A. Statistical methods for research workers. (4th ed.) Edenburgh: Oliver and Boyd, 1932.

Forsyth, R. Some empirical results related to the stability of performance indicators in Dyer's Student Change Model of an educational system. Journal of Educational Measurement, 1973, 10, 7-12.

Friedman, D. The use of pattern analysis for the prediction of achievement criteria using multiple performance measures. Educational and Psychological Measurement, 1972, 32, 1051-1054.

Gigliotti, R. J. The expectation pattern: an analysis of elementary school social environments. Unpublished doctoral dissertation, Michigan State University, 1972.

Glass, G. V. Response to Traub's "Notes on the reliability of residual change scores". *Journal of Educational Measurement*, 1968, *5*, 265-267.

Gordon, R. A. Issues in multiple regression. *American Journal of Sociology*, 1968, *73*, 592-616.

Guthrie, J. W. A survey of school effectiveness studies. In *Do teachers make a difference?* Washington, D. C.: U. S. Office of Education, 1970.

Hanushek, F. The production of education, teacher quality, and efficiency. In *Do teachers make a difference?* Washington, D. C.: U. S. Office of Education, 1970,

Hanushek, E. *Education and Race*. Lexington, Mass.: Lexington Books (div. of Heath and Co.), 1972.

Harris, D., Bisbee, C., and Evans, S. Further comments - misuse of analysis of covariance. *Psychological Bulletin*, 1971, *75*, 220-222.

Herriot, R. and Muse, D. Methodological issues in the study of school effects. In Kerlinger, F. (Ed.), *Review of research in education, 1*. Itasca, Ill.: Peacock, 1973.

Hilton, T. L. and Patrick, C. Cross-sectional versus longitudinal data: an empirical comparison of mean differences on academic growth. *Journal of Educational Measurement*, 1970, *7*, 15-24.

Houston, S., Duff, W., and Roy, M. Judgment analysis as a technique for evaluating school effectiveness. *Journal of Experimental Education*, 1972, *40*(4), 56-61.

Innes, T. The prediction of achievement means of schools from non-school factors through criterion scaling. Paper presented at the Southeastern Invitational Conference on Testing, Athens, Georgia, 1972.

Kerlinger, F. (Ed.) *Review of research in education, 1*. Itasca, Ill.: Peacock, 1973.

Lavin, D. E. *The prediction of academic success*. New York: Russell Sage Foundation, 1965.

Levin, H. M. A new model of school effectiveness. In *Do teachers make a difference?* Washington, D. C.: U. S. Office of Education, 1970.

Linn, R. and Werts, C. Assumptions in making causal inferences from part correlations, partial correlations and partial regression coefficients. Psychological Bulletin, 1969, 72, 307-310.

Linn, R., Werts, C., and Tucker, L. The interpretation of regression coefficients in a school effects model. Educational and Psychological Measurement, 1971, 31, 85-93.

Lohnes, P. R. Statistical descriptors of school classes. American Educational Research Journal, 1972, 9, 547-555.

Lord, F. Elementary models for measuring change. In Harris, C. W. (Ed.), Problems in measuring change. Madison, Wis.: University of Wisconsin Press, 1963.

Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68, 304-305.

Lord, F. M. Statistical adjustments when comparing pre-existing groups. Psychological Bulletin, 1969, 72, 336-337.

McNemar, Q. Psychological statistics. (4th ed.) New York: Wiley, 1969.

Marco, G. L. A comparison of selected school effectiveness measures based on longitudinal data. (Research Bulletin RB-73-20) Princeton, N. J.: Educational Testing Service, 1973.

Marks, E. and Martin, C. Further comments relating to the measurement of change. American Educational Research Journal, 1973, 10, 171-191.

Mayeske, G. Teacher attributes and school achievement. In Do teachers make a difference. Washington, D. C.: U. S. Office of Education, 1970.

Mayeske, G., Wisler, C., Beaton, A., et al. A study of our nations schools. Washington, D. C.: U. S. Government Printing Office, 1969.

Metfessel, N. and Michael, W. A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. Educational and Psychological Measurement, 1967, 27, 931-943.

Michelson, S. The association of teacher resourceness with children's characteristics. In Do teachers make a difference? Washington, D. C.: U. S. Office of Education, 1970.

Mood. A. M. Partitioning variance in multiple regression analyses as a tool for developing learning models. American Educational Research Journal, 1971, 8, 191-202.

Mosteller, F. and Moynihan, D. P. On equality of educational opportunity. New York: Random House, 1972.

Newton, R. G. and Spurrell, D. J. A development of multiple regression for the analysis of routine data. Applied Statistics, 1967, 16, 51-64.

Nichols, R. Effects of various college characteristics on student aptitude test scores. Journal of Educational Psychology, 1964, 55, 45-54.

O'Connor, E. F. Extending classical test theory to the measurement of change. Review of Educational Research, 1972, 42, 73-97.

Richards, J. M. A simple analytic model for college effects. School Review, 1966, 74, 380-392.

Rock, D., Baird, L., and Linn, R. Interaction between college effects and students' aptitudes. American Educational Research Journal, 1972, 9, 149-161.

Rosa, N. M. The recognition of regression effects as a problem in measuring achievement gains in performance contract evaluations: a proposed method to avoid the problem. Unpublished doctoral dissertation, University of Connecticut, 1972.

Spady, W. The impact of school resources on students. In Kerlinger, F. (Ed.), Review of research in education, 1. Itasca, Ill.: Peacock, 1973.

Sprott, D. A. Notes on Evans and Anastasio on the analysis of covariance. Psychological Bulletin, 1970, 73, 303-306.

Stanley, J. General and special formulas for reliability of differences. Journal of Educational Measurement, 1967, 4, 249-252.

Stephenson, R. and Beard, J. Common dimension of the school, social and economic environment in Florida: an empirical study. Florida Journal of Educational Research, 1971, 13, 49-57.

Tatsuoka, M. M. Multivariate analysis: techniques for educational and psychological research. New York: Wiley, 1971.

Tatsuoka, M. M. Multivariate analysis in educational research. In Kerlinger, F. (Ed.), Review of research in education, 1. Itasca, Ill.: Peacock, 1973.

Thorndike, E. L. The influence of chance imperfections of measures upon the relationship of initial score to gain or loss. Journal of Experimental Psychology, 1924, 7, 225-232.

Traub, R. A note on the reliability of residual change scores. Journal of Educational Measurement, 1967, 4, 253-256.

Tuckman, H. P. High school inputs and their contribution to school performance. Journal of Human Resources, 1971, 6, 490-509.

Tukey, J. W. Causation, regression, and path analysis. In Kempthorne, O., et al. (Eds.), Statistics and mathematics in biology. Ames: Iowa State College Press, 1954.

United States Department of Health, Education, and Welfare. Do teachers make a difference? Washington, D. C.: U. S. Government Printing Office, 1970.

United States Department of Health, Education, and Welfare. School achievement of children by demographic and socioeconomic factors. Washington, D. C.: U. S. Government Printing Office, 1972.

Van de Geer, J. P. Introduction to multivariate analysis for the social sciences. San Francisco: W. H. Freeman, 1971.

Walberg, H. Generalized regression models in educational research. American Educational Research Journal, 1971, 8, 71-91

Ward, J. H. Partitioning of variance and contribution or importance of a variable: a visit to a graduate siminar. American Educational Research Journal, 1969, 6, 467-474.

Webster, H. and Bereiter, C. The reliability of changes measured by mental test scores. In Harris, C. W. (Ed.), Problems in measuring change. Madison, Wis.: University of Wisconsin Press, 1963.

Welsh, J. Educational quality assessment: Pennsylvania looks at its schools. Harrisburg, Pa.: Pennsylvania State Department of Education, 1971.

Werts, C. E. The partitioning of variance in school effects studies. American Educational Research Journal, 1968, 5, 311-317.

Werts, C. E. The partitioning of variance in school effects studies: a reconsideration. American Educational Research Journal, 1970, 7, 127-132.

Werts, C. E. and Linn, R. L. Analyzing school effects: how to use the same data to support different hypotheses. American Educational Research Journal, 1969, 6, 439-447.

Werts, C. E. and Linn, R. L. A general linear model for studying growth. Psychological Bulletin, 1970, 73, 17-22.

Werts, C. E. and Linn, R. L. Analyzing school effects: ANCOVA with a falible covariate. Educational and Psychological Measurement, 1971, 31, 95-104. (a)

Werts, C. E. and Linn, R. L. Considerations when making inferences within the analysis of covariance model. *Educational and Psychological Measurement*, 1971, *31*, 407-416. (b)

Werts, C. E. and Linn, R. L. Problems with inferring treatment effects with repeated measures. *Educational and Psychological Measurement*, 1971, *31*, 857-866. (c)

Werts, C. and Watley, D. Analyzing college effects: correlation vs. regression. *American Educational Research Journal*, 1968, *5*, 585-598.

Werts, C. and Watley, D. A student's dilemma: big fish - little pond or little fish - big pond. *Journal of Counseling Psychology*, 1969, *16*, 14-19.

Winer, B. J. *Statistical principles in experimental design*. (2nd ed.) New York: McGraw-Hill, 1971.

Worthington, L. and Grant, C. Factors of academic success: a multivariate analysis. *Journal of Educational Research*, 1971, *65*, 7-10

Yap, K. O. A study of the efficiency of causal analysis conducted on panel data. Unpublished doctoral dissertation, University of Hawaii, 1973.